Archiving in 2050

Looking to the future

Geert-Jan van Bussel

How do we archive in the future? How will information be stored and accessed? What role will automation and AI play? Nobody knows what the future will bring, but with today's knowledge we can sketch a future vision of archiving in 2050.

In 2012, in 'Archiving should be just like an Apple', I summarized how archiving could evolve, namely creating an environment in which archiving is fully automated. [1] The idea was that users would have to do little or no manual work to process and store information contextually and that archiving would become a background process.

The manner in which information is stored and archived should not be a concern for people in organizations. They need access to information at any time, with the added benefit of metadata which enables them to assess the context, relevance and quality of the information in question. The same can be said of information that is preserved as cultural heritage in archival repositories. It is essential that users of this information are able to utilize it in an intuitive manner, locating and employing the archival information with the full context required for its interpretation. Such processes can be supported by a variety of tools and techniques, including artificial intelligence (AI) and machine learning (ML). Information should be readily accessible at the user's discretion, accompanied by a comprehensive context of its creation, provenance, utilization, and management.

In the book 'De Informatiemaatschappij van 2023' ('The Information Society of 2023'), which I edited in 2013, a number of researchers, entrepreneurs and practitioners presented their views on the irreversible trends that they believed would characterize the information society of 2023. Many trends were discussed, some of which have not been realized, while others have been. While the notion of a 'universe of loose sand' has yet to materialize, the phenomenon of 'choking on data' has emerged as a significant concern. The digitization of the world continues apace, with information becoming both social and political in nature. I identified six trends that appeared to be nascent but had not yet reached fruition. It can be argued that this is not the case at the present time either. The transition from supply-driven to demand-driven tools is still underway. An increasing amount of information is being made available to match personal profiles and search queries. It is, however, questionable whether this is always the most appropriate information. The information society is contextual, yet a considerable amount of information is stored and retained in a manner that is incongruous with its original context. The concept of 'linked data' has become a ubiquitous feature of modern life, with individuals frequently accessing information via mobile devices. However, this data is often presented in an incomplete and context-specific manner. Furthermore, the 'open' information society has been found to be everything but 'open' and transparent. Two developments that I considered to be defining trends were not referenced in the contributions. These were the environmental impact of information technology and the restrictive role of legislation. Both of these trends have intensified in recent years. The effects of climate change mean

[1] G.J. van Bussel (2012). *Archiving should be just like an Apple, en acht andere (nuttige ?) stellingen,* Amsterdam University Press, HvA Publicaties, Amsterdam, 55 pp. p. 11.

that information infrastructures must be made 'green'. The effects of legislation, such as the GDPR, limit and reinforce transparency, accessibility and openness of information. [2]

The future is inherently uncertain. However, based on the current state of information technology, it is possible to project what the year 2050 might bring. This is a sufficiently distant future to allow for discrepancies between my expectations and subsequent reality, which is a possibility that is, in fact, high.

My expectations for 2050 are based on the following assumptions:

1. I consider information to be data and data objects. What qualifies for archiving are 'reproducible collections of coherent data, carried, displayed or correlated as a unit using a form of presentation by means of a medium, with the intention of conveying information'. [3] Data may be stored in multiple places, provided that the collections of which they are a part (database records, documents, GIS datasets, CAD drawings and so on) can be reconstructed, presented and made available as they once were at an earlier point in time.
2. The concept of 'preserving information at the source' has been a long-standing idea in the field of information management. Over time, the interpretation of the term 'source' has evolved, with different interpretations emerging over the past two decades. In this context, I define 'source' as the entity responsible for generating (and receiving) the information in question, namely the organization that archives this information. I apply this definition consistently across dynamic and static forms of organizational information. It is not limited to the dynamic datasets that may be subject to a 'transfer waiver', as has been proposed in discussions about the new archives law.
3. My perspective encompasses the entirety of the information value chain, spanning the entire spectrum from generation to destruction and/or final preservation, and extending to the public consultation of information by researchers (ranging from genealogists to academics). This perspective is not constrained by the specific repository of the information in question, nor by the means of its access.
4. This perspective is primarily concerned with government organizations. While the concept is applicable to business organizations, its specific interpretation differs. However, the underlying principle remains consistent.

In 2050:

A.
Regardless of their location, employees log in to their employer's dashboard portal with their mobile work units, which are fully encrypted with post-quantum cryptography. Following authentication via both face and iris scan, they are presented with a personal dashboard that is compiled based on their authorizations.

B.
This personal dashboard shows an overview of work in progress, mostly decisions for which the organization has defined in its process definitions that human decisions are required. The staff member's work consists of reviewing the AI's daily case processing reports, checking for discrepancies, consulting with affected citizens and businesses, evaluating selected strategies and communicating any changes to the Al avatar for further processing, and providing information to managers and directors on politically

[2] G.J. van Bussel (ed.) (2013). *De informatiemaatschappij van 2023. Perspectieven op de nabije toekomst,* Lectoraat DA&C/ GEA Consultants, Amsterdam.
[3] G.J. van Bussel (2008). *Documenten onder controle. Optimaliseer uw informatievoorziening*, Deventer, p. 18 (translated).

sensitive matters. The Al avatar is always available for support and is equipped with advanced speech technology.

C.
The majority of business processes are conducted entirely automatically in accordance with Al-powered 'business rules' within applications that have been designed to handle specific and complex processes. In addition, the involvement of multiple organizations in these processes is possible, and communication between apps is enabled through the use of secure connections. The Al system utilizes knowledge bases, wherein all specifications, requirements and context data are defined per process and process step and assigned automatically during process handling. Each instance of non-compliance with the prescribed workflows is allocated to three human reviewers (potentially across multiple organizational boundaries), who each make an independent determination. The Al, in conjunction with Al units from other organizations, determines the potential consequences of the responses and the rationale provided for them. It then decides on the appropriate course of action, discusses this with the reviewers and, if permitted, incorporates the deviation into the existing definitions automatically. The ML system ensures that the algorithms of the local Al and the apps used are continuously improved in their ability to handle complex processes. The Al offers the adaptation or addition of business rules and the reasons for this to the central knowledge base (of federal, regional, and local governments) to adapt the standard models (where necessary). Policy and consultation processes are given automatic workflows, with policy co-workers being assisted by the Al. The Al avatar can, in discussion with employees, adjust the procedure to be followed and add (internal and external) participants and provide them with the requisite authorizations.

D.
The central knowledge bases contain all standardized data models of process definitions, metadata, business rules, retention schedules, archival models for long-term access, and so forth, which are used in the local knowledge bases as guidelines for an organization's specific data models. The central knowledge bases are utilized by local AI to ensure the currency of the catalogue models, to evaluate the legitimacy of local adaptations to process definitions and to ascertain whether decisions on deviations are permissible in their consequences.

E.
Each organization has an infrastructure secured with quantum technology, based on a Hybrid Storage Area Network (25th generation, $HSAN_{25}$), in which all provisions have been made for secure storage and communication of data and data objects. $HSAN_{25}$ is both on-premises and in the cloud, with the ICT branch of the central organizations providing the necessary facilitation. In accordance with the central guidelines, the Al is responsible for determining the location and type of information stored. This necessitates the involvement of multiple commercial entities, acting as suppliers of the $HSAN_{25}$ and data centres in which hosting is conducted. The data centres operate on a combination of renewable energy sources, with hydrogen representing the primary source, while solar, wind and water also contribute. All storage is redundant and equipped with two synchronizations and three mirrors, which guarantee throughput in case of emergencies. In the $HSAN_{25}$, the 'Digital Vault' has also been realized, which serves for that information that must be permanently preserved. In consideration of the climate crisis, the five data centres used are located in regions of Europe that are not susceptible to significant sea level rise.

F.
Through the personal dashboard, each employee can find all the information he or she needs to do his or her job, including all data about the context of the information and possible relationships with other

(available) information. The Al avatar can provide or suggest further information in consultation with the employee and taking into account his or her permissions. Where appropriate, the avatar may also provide answers.

G.
The management of information is largely automated, with processes aligned to the stages of the information value chain. Information is furnished with context from the initial intake stage through to the final storage stage, or until it is destroyed. Only data pertaining to individuals that is pertinent to the matter in hand is recorded; this is accessible only to those employees who are duly authorized to view this personal information. The implementation of archiving processes is realized. The deadline for destruction is contingent upon the outcome of the process and thus determines the precise moment of destruction. Where feasible, file formats are converted to sustainable formats (defined in format libraries as Pronom2050). The destruction and transfer processes are automated, with the archivist serving as the sole arbiter of exceptions, based on input from the Al-avatar regarding potential hotspots. The automated transfer to the digital vault is provided with the requisite formats and context for publication. The archivist is granted the requisite management authorisations. Information management is subject to external audit on a biannual basis. The audit is conducted in accordance with the knowledge base of the Central Inspectorate for Information Management of the Ministry of Information and Information Infrastructure, as prescribed by the Information Act (2039). The 'older' archives have been digitized. Any heritage material dating from before 1850 is retained in its original form and stored in central archive repositories in accordance with the principles of good management practice. This material is accessible to researchers by appointment. The entirety of this material is accessible in digital format via the Digital Vault.

H.
The Digital Vault serves as the repository for all information to be stored, accompanied by the requisite contextual data. The aforementioned vault is accessible to employees within the organization. The HSAN25 provides optimal security for the Digital Vault. A publicly accessible mirror of the vault is also provided. The mirror is searchable and furnished with an Al-avatar, which welcomes visitors and assists them in locating the desired information. The aforementioned mirror can be accessed from the organization's website, as well as from generic and public heritage sites. The information is presented in its original form, with emulations provided where necessary, accompanied by transcriptions and translations into modern Dutch generated by tools such as Transkribus6 or Translate20. Search engines have optical character recognition (OCR) indexes of all archives (including context), thereby enabling users to access information from multiple archives simultaneously.

This perspective could be partially realized now, in its basic form, based on the capabilities of current technology. This perspective implies that the concept of separate bodies for the preservation of archives is no longer tenable. The consequence of this interpretation of 'preservation at the source' is that the existing archival system will be unable to survive the possibilities offered by technological advancement. However, the aforementioned perspective remains largely unchanged when one assumes a position from within the existing system. This implies a different design of the process and infrastructure landscape, but it remains intact in terms of vision. The management and archiving of information are moving towards full automation, and AI will play an important role in this.

As with any perspective, there is a possibility that I may be mistaken. Even then, the world continues to function. ...